8:45 - 9:00 Welcome Remarks

***9:00 - 10:00 Keynote Speech:* Feedback on Feedback in Autonomic Computing Systems**

By Jeffrey Kephart, IBM Research

(abstract) From the earliest days of autonomic computing, feedback control techniques have provided powerful means for achieving desired behavior in computer systems. In his keynote at the International Conference on Autonomic Computing in 2009, Joe Hellerstein described several applications of feedback control, including allocating memory and throttling background utilities in commercial database management systems and managing thread pools. Through the years, audiences at ICAC and other venues such as the FeBID (Feedback Control Implementation and Design in Computing Systems and Networks) workshop have heard about a diverse set of feedback control applications ranging from power management in servers and data centers to management of virtualized resources. During this time, the notion of what it means to apply feedback principles to computing has been broadened considerably - an evolution that is reflected in this year's renaming of FeBID to "Feedback Computing". What does "feedback computing" really mean? Is this broader construal useful, or have we lost our focus? I will offer some opinions, based on examples of recent work from my research lab and from other institutions, and will conclude by seeking feedback from the audience.

10:00 - 10:30 Coffee Break

10:30 - 12:00 Technical Session 1: Control Theory and Applications
Chair: Eric Rutten

Overload Management in Data Stream Processing Systems with Latency Guarantees
Evangelia Kalyvianaki (Imperial College London)
Themistoklis Charalambous (Royal Institute of Technology (KTH))
Marco Fiscato (Imperial College London)
Peter Pietzuch (Imperial College London)
(abstract) Stream processing systems are becoming increasingly important to analyse real-time data generated by modern applications such as online social networks. Their main characteristic is to produce a continuous stream of fresh results as new data are being generated at real-time. Resource provisioning of stream processing systems is difficult due to time-varying workload data that induce unknown resource demands over time. Despite the development of scalable stream processing systems, which aim to provision for workload variations, there still exist cases where such systems face transient resource shortages. During overload, there is a lack of resources to process all incoming data in real-time; data accumulate in memory and their processing latency grows uncontrollably compromising the freshness of stream processing results. In this paper, we present a feedback control approach to design a nonlinear discrete-time controller that has no knowledge of the system to be controlled or the workload for the data and is still able to control the average tuple end-to-end latency in a single-node stream processing system. The results, of our evaluation on a prototype stream processing system, show that our method controls the average tuple end-to-end latency despite the time-varying workload demands and increasing number of queries.

A Model-based Framework for Automatic Recovery from Incipient Faults in Computing Systems
Rui Jia (Mississippi State University)
Sherif Abdelwahed (Mississippi State University)
Abdelkarim Erradi (Qatar University)
Rashid Hadjidj (Qatar University)
Asim Ali (Qatar University)
(abstract) This paper introduces a model-based fault management framework for computing systems. The proposed

approach targets incipient faults in which available system measurements can be used for real-time fault detection and identification. The proposed framework integrates control and diagnosis processes within a model-based framework that enables computing systems to adapt themselves to variations in their environment, maintains the system availability and reliability, and facilitates system recovery from failures while minimizing the impact on performance. The paper presents the main components of the fault adaptive framework and introduces the underlying mathematical formulation of the management problem. A case study of a memory leak fault is presented to demonstrate the framework. We also discuss the potential application of this framework to other types of failures and its extension to distributed systems.

Towards a Passivity Framework for Power Control and Response Time Management in Cloud Computing

Michael D. Lemmon (University of Notre Dame)

(abstract) There has been great interest in using classical control theory to manage computing systems. Classical control, however, focuses on regulating a system's state in a neighborhood of an equilibrium point and it is unclear if such equilibrium based methods are well-suited for systems providing performance guarantees in the face of large and rapid input fluctuations. This may be the case for cloud computing applications where consumer workloads vary in a rapid and unpredictable manner. As an alternative to classical methods, this paper discusses a passivity framework for power control and response-time management in cloud computing applications. This paper suggests that passivity concepts provide a decentralized method for certifying whether a collection of interconnected cloud computing systems can coordinate their actions in a stable manner.

Event-Based Response Time Estimation

Manfred Dellkrantz (Lund University)
Maria Kihl (Lund University)
Anders Robertsson (Lund University)
Karl Johan Åström (Lund University)

(abstract) Response time is a measure of quality of service in computer systems. Estimation techniques, suitable for support systems for mobile phone systems, are explored. These systems are complex queueing systems with large databases. The trace generated by users and system administrators changes rapidly, some loads can be measured other cannot. Attempts to capture all details give models that are not suitable for on-line control. Estimators based on continuous flow models with event based measurements are designed using extended Kalman Filtering. The estimators are compared with simple-data based estimators.

12:00 - 1:00 Lunch

1:00 - 2:00 Group Discussion

2:00 - 3:00 Technical Session 2: Performance Optimization

Chair: Sharad Singhal

vATM: vSphere Adaptive Task Management

Zhichao Li (Stony Brook University)
Aalap Desai (VMware)
Chirag Bhatt (VMware)
Rajit Kambo (VMware)
Erez Zadok (Stony Brook University)

(abstract) Virtualization provides several benefits to users in the datacenter in terms of infrastructure cost savings (e.g., capital, power, space, cooling, labor). Examples include highly efficient and available resource, networking, and storage management. As many workloads have moved to virtualized environments, it is critical that vSphere handles scale and performs optimally. However, there are scenarios where the vSphere platform exhibits suboptimal performance if user-level operations are scheduled poorly. In this paper we propose a feedback-based approach to maximize the platform performance of vSphere with a gradient based hill climbing algorithm. We have implemented the gradient based hill climbing approach. Our initial results show promising performance improvements, in terms of end-to-end latency, under the vSphere environment.

Intermediate Deadline Assignment for Distributed Real-Time Systems: Utility Maximization and Challenges

Jinkyu Lee (University of Michigan)
Insik Shin (KAIST)

(abstract) It is generally difficult to determine end-to-end delays of distributed applications such that the aggregate system utility is maximized. Many existing approaches suggest the use of intermediate deadlines of tasks in order to control and upper-bound their end-to-end delays. This paper recapitulates a utility maximization problem formulation using intermediate deadlines and its distributed solutions using price signals, both of which have been proposed in our previous work. Then, we discuss two interesting issues regarding the utility maximization problem. First, we design utility functions that adjust a tradeoff between performance (in terms of minimizing total delays) and delay fairness. Second, we derive a tighter problem formulation by using a necessary and sufficient real-time schedulability analysis technique and discuss its challenges. Since it is difficult to find an optimal solution of the tighter formulation due to non-convexity, we consider to apply feedback control techniques. Regarding this, we introduce our ongoing idea with challenges.

Data Centers as Demand Response Resources in the Electricity Market: Some Preliminary Results

Rui Wang (Drexel University)
Nagarajan Kandasamy (Drexel University)
Chika Nwankpa (Drexel University)

(abstract) Electric utilities have recently instituted demand response (DR) programs, with economic incentives that encourage consumers to modify their demand level and usage patterns during periods of peak load as well as grid emergencies. Data centers, being major consumers of power, can play an important role in the efficient operation of electrical grids. This paper describes work aimed at developing an optimization framework that allows data centers to be treated as DR resources that can effectively participate in wholesale energy markets by reducing the consumption of electricity in response to signals from utility companies. As a return, data centers are paid a reward based on the prevailing market price of electricity. We use a case study involving a set of geographically distributed data centers participating in an economic DR program to validate the framework.

3:00 - 4:00 Poster Session (and Coffee Break)

Towards Feedback-Based Generation of Hardware Characteristics

Marcus Jägemar (Ericsson AB)
Sigrid Eldh (Ericsson AB)
Andreas Ermedahl (Ericsson AB)
Björn Lisper (Mälardalen University)

(abstract) In large complex server-like computer systems it is difficult to characterise hardware usage in early stages of system development. Many times the applications running on the platform are not ready at the time of platform deployment leading to postponed metrics measurement. In our study we seek answers to the questions: (1) Can we use a feedback-based control system to create a characteristics model of a real production system? (2) Can such a model be sufficiently accurate to detect characteristics changes instead of executing the production application? The model we have created runs a signalling application, similar to the production application, together with a PID-regulator generating L1 and L2 cache misses to the same extent as the production system. Our measurements indicate that we have managed to mimic a similar environment regarding cache characteristics. Additionally we have applied the model on a software update for a production system and detected characteristics changes using the model. This has later been verified on the complete production system, which in this study is a large scale telecommunication system with a substantial market share.

SMART Computing Systems: Sensing, Modelling, Actuating, Regulating, and Tuning

Martina Maggio (Lund University)
Alessandro Vittorio Papadopoulos (Politecnico di Milano)
Alberto Leva (Politecnico di Milano)

(abstract) Control-based decision mechanisms are nowadays exploited as a viable tool to obtain reliability and adaptivity in complex computing systems. However, due to the complexity of these systems, a lot of effort is often devoted to configuring out what strategy is the best fit for the problem. This leads to ad hoc solutions that are perfect for a specific issue but are not keen to be reused. In this position paper, some reasoning is presented on one approach that can be followed when dealing with a control problem in complex computing systems. The framework is named SMART, which stands for Sensing, Modeling, Actuating, Regulating and Tuning. The considerations

discussed herein come from experiences with feedback scheduling and feedback “application-aware" resource allocation. Relevant case studies are referenced for the interested reader.

Two-Stage Prediction for CPU Time Allocation in Soft Real-Time Systems
Safayet N Ahmed (Georgia Institute of Technology)
Bonnie Ferri (Georgia Institute of Technology)
(abstract) This paper presents a prediction algorithm for CPU time allocation in soft real-time systems. The algorithm consists of two stages. The first stage predicts computation times by exploiting correlation between the computation times of successive jobs. The second stage estimates the prediction error based on the mean and standard deviation of previous prediction errors. It is shown that the proposed algorithm is able to guarantee bounds on deadline miss-rates without imposing constraints or assuming knowledge of the distribution of computation times. Three different first-stage predictors are explored. Simulation results are presented for MPEG4 video decoding times.

A Hybrid System Approach to Model Dynamic Information Flow Tracking
Maria Khater (University of New Mexico)
Rafael Fierro (University of New Mexico)
Antonio Espinoza (University of New Mexico)
Jedidiah R. Crandall (University of New Mexico)
*(abstract)*Control theory is used in several disciplines such as mechanical engineering, communication, and computer engineering. This paper examines the application of control theory to computer security, specifically in the area of Dynamic In- formation Flow Tracking (DIFT). Control theory can be used to model the destruction of information in DIFT. The DIFT system examined in this paper utilizes a taint vector to record the program's tainted data, however, as with any DIFT system, there are several cases which cause undesired results such as over-tainting. This incurs an unacceptable loss of information. This problem is solved by designing a hybrid controller which guarantees that the system will not over-taint, while still allowing taint information to flow properly. The model accomplishes this by tracking the dynamics of the program at the assembly level. Therefore the proposed controller must be included in the design in order to solve the problem traditional methods have yet to properly address. This paper provides a new approach to modeling the destruction of information in DIFT. In addition, it examines a novel application of hybrid control systems in computer security.

4:00 - 5:00 Technical Session 3: Learning and Coordination
Chair: Sherif Abdelwahed

On How to Coordinate the Behavior of Independent Adaptive Systems
Jacopo Panerati (Politecnico di Milano)
Marco Triverio (Politecnico di Milano)
Martina Maggio (Lund University)
Marco Domenico Santambrogio (Politecnico di Milano)
(abstract) Nowadays the complexity of computing systems is skyrocketing. Programmers have to deal with extremely powerful computing systems that take time and considerable skills to be instructed to perform at their best. This work analyzes the stated problem and proposes a simple, yet powerful mechanism for optimizing performance through the coordination of the interaction of multiple, independent adaptive systems called services. In this scenario we developed the Service Coordinator, a system-centralized decision engine based on reinforcement learning. The Service Coordinator gathers information about the performance goals of the sys- tem and it can either turn services on or off. The Service Coordinator analyzes the runtime impact of services and of their autonomous decision policies, looking for a combination of services that makes it possible to reach the given goals. The experiments that have been carried out show the ability of the Service Coordinator to adapt to changing conditions, confirming the validity and the flexibility of the followed approach.

Modeling VM Performance Interference with Fuzzy MIMO Model
Lixi Wang (Florida International University)
Jing Xu (Florida International University)
Ming Zhao (Florida International University)

(abstract) Virtual machines (VM) can be a powerful platform for multiplexing resources for applications workloads on demand in datacenters and cloud systems. However, it remains challenging for the resource management in such a virtualized system to deliver performance guarantees because of the contention on non-partitionable sources such

as last-level CPU cache, memory bandwidth, and on-disk buffer, which introduces performance interference between co-hosted VMs and cannot be directly controlled. To address this challenge, this paper proposes to use fuzzy modeling to establish a multi-input-multi-output performance model for co-hosted VMs in order to capture their coupling. Based on the model, the level of contention on the competing resources is quantified by the model parameters, and can be used to guide VM placement and resource allocation decisions. Experimental evaluations on application benchmarks TPC-H and RUBiS demonstrate that this fuzzy modeling approach can detect and quantify the interference from VMs competing for both CPU and I/O; it can better capture the variation of such contention compared to a linear model.

Discrete control for the coordination of administration loops (extended abstract)

Soguy Mak-Karé Gueye (LIG / UJF)
Noël De Palma (LIG / UJF)
Eric Rutten (LIG / INRIA)

(abstract) The increasing complexity of computer systems has led to the automation of administration functions, in the form of autonomic managers. One important aspect requiring such management is the issue of energy consumption of computing systems, in the perspective of green computing. As these managers address each a specific aspect, there is a need for using several managers to cover all the domains of administration. However, coordinating them is necessary for proper and effective global administration. Such coordination is a problem of synchronization and logical control of administration operations that can be applied by autonomous man- agers on the managed system at a given time in response to events observed on the state of this system. We therefore propose to investigate the use of reactive models with events and states, and discrete control techniques to solve this problem. In this paper, we illustrate this approach by integrating a controller obtained by synchronous programming, based on Discrete Controller Synthesis, in an autonomic system administration infrastructure. The role of this controller is to orchestrate the execution of reconfiguration operations of all administration policies to satisfy properties of logical consistency. We apply this approach to coordinate three managers: two energy-aware ones, which control server pro-visioning and processor frequency, and a repair manager.